

Moderator: Heidi Chumley, MD

Discussant: Maxine Papadakis, MD

## The Professionalism Mini-Evaluation Exercise: A Preliminary Investigation

Richard Cruess, Jodi Herold McIlroy, Sylvia Cruess, Shiphra Ginsburg, and Yvonne Steinert

### Abstract

#### Background

As the evaluation of professional behaviors has been identified as an area for development, the Professionalism Mini-Evaluation Exercise (P-MEX) was developed using the mini-Clinical Examination Exercise (mini-CEX) format.

#### Method

From a set of 142 observable behaviors reflective of professionalism identified at a McGill workshop, 24 were converted into an evaluation instrument modeled on the mini-CEX. This instrument,

designed for use in multiple settings, was tested on clinical clerks in medicine, surgery, obstetrics and gynecology, psychiatry, and pediatrics. In all, 211 forms were completed on 74 students by 47 evaluators.

#### Results

Results indicate content and construct validity. Exploratory factor analysis yielded 4 factors: doctor-patient relationship skills, reflective skills, time management, and interprofessional relationship skills. A decision study

showed confidence intervals sufficiently narrow for many measurement purposes with as few as 8 observations. Four items frequently marked below expectations may be identifiers for "problem" students.

#### Conclusion

This preliminary study suggests that the P-MEX is a feasible format for evaluating professionalism in clinical training.

*Acad Med.* 2006;81(10 Suppl):S74-S78.

Transmitting professionalism and professional values by role modeling is no longer recognized as sufficient.<sup>1</sup> Most faculties of medicine have established programs to explicitly teach professionalism and improve role modeling. As this occurred, it became apparent that the evaluation of professionalism needed improvement and there have been calls for new methods to be developed.<sup>2-8</sup> These have been given added urgency by studies indicating that lapses in professional behavior observed in medical school are associated with subsequent unprofessional conduct in practice.<sup>9,10</sup>

Norcini and colleagues reported on the use of the mini-Clinical Examination Exercise (mini-CEX) for residents in internal medicine,<sup>11-13</sup> noting that the results correlated well with those of other performance-based assessments.<sup>14</sup> The format is easy to use, promotes feedback, and its reliability and validity have been confirmed,<sup>15</sup> with good interexaminer reliability.<sup>16</sup> It was also shown to be feasible, reliable, and valid in

summative<sup>17</sup> and formative assessment with clinical clerks in internal medicine.<sup>18</sup>

Although the mini-CEX includes "humanistic qualities/professionalism" as one category, it does not identify specific behaviors to be observed in realistic contexts, a prerequisite for evaluating professionalism.<sup>4</sup> Because the format is so attractive, we decided to develop a tool based upon the mini-CEX to evaluate professional behaviors in medical students, and have called it the Professionalism Mini-Evaluation Exercise (P-MEX).

#### Method

##### Instrument development

As part of an ongoing faculty development program on teaching and evaluating professionalism,<sup>19</sup> a workshop on evaluating the attributes of the physician was attended by 92 McGill faculty members and residents. A consensus was achieved on the attributes of the professional and the behaviors which reflect these attributes. The workshop utilized a published definition of profession,<sup>23</sup> a definition of the healer,<sup>24</sup> and a list of the attributes of the healer and the professional drawn from

the literature<sup>25-29</sup> and used in teaching at McGill. Small groups were asked to describe behaviors consonant with the attributes and then consider methods of evaluating each behavior. The participants identified 142 behaviors similar to those developed by the National Board of Medical Examiners/ Association of American Medical Colleges workshop<sup>30</sup> and by colleagues in evaluating surgical residents.<sup>31</sup> The authors subsequently distilled them to 24 behaviors to evaluate as many attributes as possible with the smallest number of behaviors.

The selected behaviors were then inserted into the mini-CEX format using a 4-point scale where 4 = exceeded expectations, 3 = met expectations, 2 = below expectations, and 1 = unacceptable. There was also a fifth category entitled "not observed" or "not applicable." As the form is intended to be utilized in situations with and without patients, it was anticipated that this last category would be used when patient-related behaviors were not observable. Instructions for use were printed on the form, and there was space for information about the student, the

Correspondence: Richard Cruess, MD, Centre for Medical Education, Lady Meredith House, McGill University, 1110 Pine Ave. W., Montreal, QC, H3A 1A3 Canada; e-mail: (richard.cruess@mcgill.ca).

evaluator, the service and setting in which the evaluation took place, for comments, and for reporting critical incidents.

The P-MEX was designed to be used in any situation where a student's behavior can be observed, including patient encounters, small group sessions, and sign-out rounds. The evaluation is to be based on interactions that are relatively short and that occur frequently as a part of training so that each student can be evaluated on several occasions by different faculty members. Each form has two copies, one of which is given to the student, the other being retained. The evaluator is expected to give timely feedback to the student, thus giving it the potential to be formative and summative.

The P-MEX was field tested in two ways. First, during a simulation exercise, 38 individuals from diverse specialties were asked to recall a recent incident involving a medical student and to use the form to evaluate the student's performance. Secondly, four physicians on a general internal medicine ward completed it after observing an encounter between a student and a patient. Based on the feedback received from the field testing, the form was revised and the pilot project started.

IRB approval was granted and informed consent was obtained from participating students and evaluators. Students, residents, and faculty were free to decline participation. Because senior residents routinely participate in student evaluation, some served as evaluators. Testing proceeded on students during third- and fourth-year clerkships in Internal Medicine, General Surgery, Pediatrics, Psychiatry, and Obstetrics and Gynecology at McGill University. Instructions on using the form were given to all evaluators, who determined when an activity would be evaluated.

The research associate conducted semistructured interviews with all students and evaluators to assess their perception of the use, benefits, and limitations of the P-MEX.

### Data analysis

Descriptive statistics were computed on demographic and contextual data, such as who was evaluating whom, and in what setting. The item analysis consisted

of frequencies of each score category, with a focus on the "below expectations" and "not applicable" options. Items with high proportions of responses in the "not applicable" category may not be particularly useful in this particular form of evaluation, whereas those items with high proportions of responses in the "below expectations" category may be the most sensitive to breaches of professionalism. In addition to frequency distributions per item, a  $24 \times 24$  correlation matrix was generated to identify highly correlated or "redundant" items for the purpose of item reduction.

Exploratory factor analysis was conducted to understand the internal structure of the scale. The analysis was conducted in SPSS using unweighted least squares extraction and varimax rotation. The factor analysis was conducted using form ( $n = 211$ ) as the unit of analysis. Each occasion of measurement (form) should "stand alone," because other sources of variance (i.e., context/occasion/rater) are all confounded and likely contribute sufficient measurement that each measurement circumstance for a given student could be considered independent of the others.

Additionally, a generalizability analysis<sup>32</sup> and decision study was performed to investigate the number of forms (occasions of measurement) required to obtain a dependable estimate of the calculated average score. This analysis was done using the approach described by Norcini and colleagues<sup>11,14</sup> in that the persons by forms (occasions) design was used, and the calculated average score (i.e., mean over the 24 items) was the score of interest. Using subjects for whom at least two forms were completed, the variance components were estimated for student, occasion, and the student by occasion interaction (error) using urGENOVA.<sup>33</sup> These variance components were then used in a series of decision studies using GENOVA to compute the reproducibility coefficients, and standard errors of measurement were computed for 1 to 14 encounters/occasions. Confidence intervals were computed around the mean calculated average score to obtain a sense of the precision of measurement in each case.

## Results

### Demographics

In all, 211 P-MEX forms were collected on 74 undergraduate medical students. The number of forms per student in the 2-rotation study period ranged from 1 to 9, with a mean of  $2.85 (\pm 1.9)$  and a median of 2.0. When we eliminate those subjects for whom there was only one form, there are 189 forms on 52 subjects, with a median number of forms per subject of 3.5. These data are provided to allow comparison with the preliminary investigation by Norcini et al.<sup>11</sup> The 211 P-MEX forms were completed by 47 evaluators; 34 (72%) were faculty and 13 (28%) senior residents. The number of forms completed per evaluator ranged from 1 to 16, with a mean of  $4.5 (\pm 3.3)$  and a median of 4.0.

The context of evaluation varied greatly. The setting was identified by the evaluator as "ward activity" (24%), "bedside rounds" (13%), "ambulatory clinic" (14%), "OR/emergency room" (10%), "sign-out rounds" (9%), "small group teaching" (8%), and "team meeting" (2%). A total of 9% of the forms were marked "other" for setting, and 12% were either blank or more than one setting was selected.

### Item analysis

On average, 6 of the 24 items were not completed or were marked "not applicable." Four items were marked not applicable in 40% or more of the 211 forms: "accepted inconvenience to meet patient needs," "advocated on behalf of a patient and/or family member," "admitted errors/omissions," and "assisted a colleague as needed." They may not be as relevant to evaluation in the P-MEX context as other items with lower frequency use of that category.

Item mean scores ranged from 3.10 to 3.35 out of 4. 4 items showed 3% or more of the ratings as "below expectations": "demonstrated awareness of limitations," "solicited feedback," "was on time," and "addressed gaps in own knowledge and skills." This may indicate that these items are more "sensitive" to breaches of professionalism than others. Three items were seen to be redundant: "showed respect for patient" (correlation with items 1 and 2,  $r = 0.78$  and  $0.79$ ), "assisted a colleague as needed" (correlation with items 22, 23 and 24,  $r =$

0.87, 0.82, and 0.83, respectively), and “respected rules and procedures of the system” (correlation with items 14, 21, 22 and 23,  $r = 0.79, 0.83, 0.77,$  and  $0.76,$  respectively).

### Scale analysis

The exploratory principal components analysis revealed four factors with eigenvalues over 1.0, accounting for 85% of the variance in the 24 items. The rotated solution shows that the factors each accounted for 13–26% of the total variance. The factor loadings of the 24 items onto the four factors are shown in Table 1. The items appear to cluster into factors which can be interpreted as: Doctor-Patient Relationship Skills (items 1 through 7, 24% of total variance); Reflective Skills (items 8 to 11 and 13,

19% of total variance); Time Management (items 15, 16 and 18, 13% of total variance); and Interprofessional Relationship Skills (items 14, 17, and 19 through 24, 26% of total variance). Note that item 12, “Maintained appropriate boundaries with patients/colleagues,” loads equally onto the Doctor-Patient Relationship Skills and the Interprofessional Relationship Skills factors, reflecting the “double-barreled” nature of the item. There were a number of other items whose factor loading coefficients were very high on one factor but also somewhat high on at least one other factor.

The average score for the 24 items was computed for each form, and then aggregated over students and over raters.

The mean computed average score on the 211 forms was 3.25, and the mean score aggregated by student was 3.24 (range 2.70 to 4.0) and aggregated by raters was 3.21 (range 2.85 to 4.0). The reproducibility of the average score (across 24 items) was estimated using generalizability theory. The universe score variance component for the calculated average score was 0.036 and the error variance component was 0.093. This error variance component can be interpreted as the within-student variation in ratings over multiple occasions of evaluation. Using these two estimates, the reproducibility coefficients and standard errors of measurement were computed for 1 to 14 encounters/occasions (Table 2). Between 10 and 12 encounters are required to obtain a reproducibility coefficient of 0.80, which is consistent with results published by Norcini and colleagues in 1995.<sup>11</sup> From the standard error of measurement, corresponding 95% confidence intervals for a student with a calculated average score of 3.25 were computed over incremental numbers of forms. These are reported in Table 2. Using these confidence intervals as guidelines, educators may feel comfortable using the average of fewer than 10 forms, depending on the intended use of the scores. For example, the confidence interval obtained with an SEM of 0.11 may be sufficiently precise for many criterion-referenced measurement purposes, requiring only eight forms per student to be collected.

Table 1

Rotated Factor Matrix Solution for Factor Analysis of 24 Items

	Factor			
	1	2	3	4
<b>Doctor-patient relationship skills</b>				
1. Listened actively to patient	.493	<b>.668</b>	.372	.175
2. Showed interest in patient as a person	.282	<b>.812</b>	.359	.219
3. Showed respect for patient	.427	<b>.752</b>	.333	.184
4. Recognized and met patient needs	.417	<b>.708</b>	.245	.260
5. Accepted inconvenience to meet patient needs	.344	<b>.677</b>	.449	.227
6. Ensured continuity of patient care	.285	<b>.750</b>	.290	.436
7. Advocated on behalf of a patient and/or family member	.307	<b>.631</b>	.298	.440
12. Maintained appropriate boundaries with patients/colleagues	<b>.528</b>	<b>.562</b>	.393	.189
<b>Reflective skills</b>				
8. Demonstrated awareness of limitations	.459	.344	<b>.632</b>	.019
9. Admitted errors/omissions	.257	.249	<b>.783</b>	.094
10. Solicited feedback	.248	.335	<b>.783</b>	.314
11. Accepted feedback	.199	.247	<b>.825</b>	.245
13. Maintained composure in a difficult situation	.444	.419	<b>.598</b>	.224
<b>Time management</b>				
15. Was on time	.248	.228	.327	<b>.804</b>
16. Completed tasks in a reliable fashion	.321	.447	.063	<b>.632</b>
18. Was available to patients or colleagues	.459	.228	.183	<b>.746</b>
<b>Interprofessional relationship skills</b>				
12. Maintained appropriate boundaries with patients/colleagues	<b>.528</b>	<b>.562</b>	.393	.189
14. Maintained appropriate appearance	<b>.648</b>	.270	.515	.304
17. Addressed own gaps in knowledge and skills	<b>.523</b>	.377	.338	.317
19. Demonstrated respect for colleagues	<b>.726</b>	.249	.186	.381
20. Avoided derogatory language	<b>.777</b>	.357	.227	.248
21. Assisted a colleague as needed	<b>.722</b>	.424	.261	.258
22. Maintained patient confidentiality	<b>.797</b>	.314	.349	.257
23. Used health resources appropriately	<b>.772</b>	.388	.305	.235
24. Respected rules and procedures of the system	<b>.709</b>	.340	.365	.336

Extraction method: unweighted least squares. Rotation method: varimax with Kaiser normalization. Rotation converged in 7 iterations.

### Qualitative feedback

Preliminary analysis of the responses to the semistructured interviews indicated that the P-MEX was useful in promoting self-reflection, awareness of the importance of professionalism in daily encounters, identifying behaviors consistent with professionalism, and teaching about this subject matter. The major limitations have been time – time to observe; time to record; and time to give feedback.

### Discussion

This preliminary study suggests that the P-MEX is a feasible format for evaluating professionalism in clerkship training. Content validity of the form is evidenced through the extensive process of item generation, and the results of this process

Table 2

**Decision Study for Calculated Mean Score on Professionalism Mini Evaluation Exercise (P-MEX)**

Number of forms	Generalizability coefficient	SEM	95% Confidence interval
1	0.28	0.31	2.64 to 3.86
2	0.44	0.22	2.82 to 3.68
4	0.61	0.15	2.96 to 3.54
6	0.70	0.12	3.01 to 3.49
8	0.76	0.11	3.03 to 3.47
10	0.79	0.10	3.05 to 3.45
12	0.82	0.09	3.07 to 3.43
14	0.84	0.08	3.09 to 3.41

were “triangulated” with similar processes conducted by other groups in North America.

We have shown evidence of construct validity of the P-MEX through factor analysis in that the 24 original items cluster into identifiable factors or facets of the construct. However, due to the somewhat low sample size and ordinal nature of the rating scale, the results of this analysis are intended only for the purpose of understanding the internal structure of the scale, as opposed to justification for computation and reporting of scores on subscales corresponding to these factors.

The reproducibility of the calculated average score was shown to be comparable to that reported in the preliminary study of the mini-CEX, in that between 10 and 12 completed forms are required to achieve a dependability coefficient of 0.80. However, confidence intervals may be sufficiently small at 4 to 6 forms for many measurement purposes.

Some minor form revisions were deemed necessary as a result of the item analysis. Specifically, the item asking about setting has been changed, three redundant items (3, 21, and 24) have been eliminated, and three other items (7, 12, and 18) were reworded to eliminate their “double-barreled” nature.

One interesting finding is that the four items marked “below expectations” closely relate to reflective skills. The extent to which individual items or groups of items are predictive of future difficulties is one of the many areas for future investigation. Currently, investigations are ongoing at McGill,

using the revised form at the postgraduate training level.

As stated earlier, feedback on the P-MEX has demonstrated its value for promoting self-reflection and awareness of the importance of professionalism. Based on this experience, the P-MEX appears to be a useful assessment method that can drive teaching and learning.

### Acknowledgments

The authors would like to acknowledge the support of the ABIM Foundation and particularly Linda Blank during the conception and implementation of this study. We would also like to acknowledge Sharon Wood-Dauphinee, who was instrumental in mapping behaviors to attributes of professionalism, Linda McHarg, who was an invaluable research associate, and the faculty members, who gave generously of their time.

### References

- 1 Cruess SR, Cruess RL. Professionalism must be taught. *BMJ*. 1997;315:1674–7.
- 2 Arnold L. Assessing professional behaviors: Yesterday, today, and tomorrow. *Acad Med*. 2002;77:502–15.
- 3 Epstein R, Hundert E. Defining and assessing professional competence. *JAMA*. 2002;287:226–35.
- 4 Ginsburg S, Regehr G, et al. Context, conflict, and resolution: A new conceptual framework for evaluating professionalism. *Acad Med*. 2007;75:S6–S11.
- 5 Lynch DC, Surdyk PM, Eiser AR. Assessing professionalism: a review of the literature. *Med Teacher*. 2004;26:366–73.
- 6 Stern DT. Practicing what we preach? An analysis of the curriculum values in medical education. *Am J Med*. 1998;104:569–75.
- 7 Veloski JJ, Fields SK, Boex JR, Blank LL. Measuring professionalism: a review of studies with instruments reported in the literature between 1982 and 2002. *Acad Med*. 2005;80:366–70.

- 8 Shrank WH, Reed VA, Jernstedt GC. Fostering professionalism in medical education: a call for improved assessment and meaningful incentives. *J Gen Int Med*. 2004;19:887–92.
- 9 Papadakis MA, Teharani A, Banach MA, et al. Disciplinary action by medical boards and prior behavior in medical school. *N Engl J Med*. 2005;353:2673–82.
- 10 Kirk LM, Blank LL. Professional behavior—a learner’s permit for licensure. *N Engl J Med*. 2005;353:2709–11.
- 11 Norcini JJ, Blank LL, Arnold GK, Kimball HR. The Mini-CEX (Clinical Evaluation Exercise): A preliminary investigation. *Ann Intern Med*. 1995;123:795–9.
- 12 Noel GL, Herbers JE, Caplow MP, et al. How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Intern Med*. 1992;117:757–65.
- 13 Kroboth FJ, Hanusa BH, Parker S, Coulehan JL. The inter-rater reliability and internal consistency of a clinical evaluation exercise. *J Gen Int Med*. 1992;117:757–65.
- 14 Norcini JJ, Blank LL, Duffy D, Fortna GS. The Mini-CEX: A method for assessing clinical skills. *Ann Intern Med*. 2003;138:476–81.
- 15 Durning S, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the Mini-Evaluation Exercise for internal medicine residency training. *Acad Med*. 2002;77:900–4.
- 16 Norcini JJ, Arnold GK, Blank LL, Kimball HR. Examiner differences in the Mini-CEX. *Adv Health Sci Educ*. 1997;2:27–33.
- 17 Kogan JR, Bellini LM, and Shea JA. Feasibility, reliability, and validity of the Mini-Clinical Evaluation Exercise (Mini-CEX) in a medicine core clerkship. *Acad Med*. 2003;78:S33–S34.
- 18 Holmboe ES, Yepes M, Williams F, Huot SJ. Feedback and the Mini Clinical Exercise. *J Gen Int Med*. 2004;19:558–61.
- 19 Steinert Y, Cruess SR, Cruess RL, Snell L. Faculty development for teaching and evaluating professionalism: from program design to curricular change. *Med Educ*. 2005;39:127–36.
- 20 Cruess RL, Cruess SR. Teaching medicine as a profession in the service of healing. *Acad Med*. 1997;72:941–52.
- 21 Cassell E. *Doctoring: the nature of primary care medicine*. New York: Oxford University Press, 1999.
- 22 Papadakis M, Osborn EHS, Cooke M, Healey K. A strategy for the detection and evaluation of unprofessional behavior in medical students. *Acad Med*. 1999;74:980–90.
- 23 Cruess SR, Johnston S, Cruess RL. Professionalism: a working definition for medical educators. *Teach Learn Med*. 2004;16:74–6.
- 24 Oxford English Dictionary, 2nd ed. Oxford: Clarendon Press, 1989.
- 25 Freidson E. *Professionalism Reborn: Theory, Prophecy, and Policy*. Cambridge: Polity Press, 1994.

- 26 Swick HM. Towards a normative definition of professionalism. *Acad Med.* 2000;75:612–6.
- 27 Rabinowitz D, Reis S, Van Raalte R, Alroy G, Ber R. Development of a physician attributes database as a resource for medical education, professionalism and student evaluation. *Med Teacher.* 2004;26:160–165.
- 28 Van de Camp K, Vernooij M, Grol R, Bottema B. How to conceptualize professionalism. *Clin Teacher.* 2004;26:696–702.
- 29 Mann KV, Ruedy J, Millar N, Andreou P. Achievement of non-cognitive goals of undergraduate medical education: perceptions of medical students, residents, faculty, and other health professionals. *Med Educ.* 2005;39:40–5.
- 30 Report of an invitational conference cosponsored by the Association of American Medical Colleges and the National Board of Medical Examiners. Embedding professionalism in medical education: assessment as a tool for implementation. Washington, DC: National Board of Medical Examiners, 2003.
- 31 Gauger PG, Gruppen LD, Minter RM, Colletti LM, Stern DT. Initial use of a novel instrument to measure professionalism in surgical residents. *Am J Surg.* 2005;189:479–87.
- 32 Brennan RL. Elements of Generalizability Theory. Iowa City, IA: American College Testing Publications, 1983.
- 33 Brennan RL. urGENOVA Version 2.1. Iowa City, IA: Iowa Testing Programs.